CHAPTER 2: TEST DEVELOPMENT, ADMINISTRATION, AND SCORING

Introduction

A major concern raised in our first evaluation report was whether it was feasible to develop a high quality exam within the time constraints specified in the legislation. Surprisingly, it was. We documented our review of the quality of the test forms that were administered in 2001 and also our review of administration, scoring, and reporting procedures used for that administration in a report submitted to the legislature earlier this year (Wise et al., Feb. 2002). In this chapter, we describe our observations to date on the development, administration, and scoring of the 2002 CAHSEE test forms.

CDE and the test development contractors learned a number of lessons from the 2001 administration of the CAHSEE, and many of these lessons were translated into improved procedures for developing, administering and scoring the test in 2002. For example, in 2001, the CAHSEE was administered during two days of testing spaced a week apart. The 2002 CAHSEE was administered over three consecutive days, allowing two days for the ELA portion. This change was a direct response to expressed concerns about the logistics and burden of the 2001 administration.

Our observations and analyses of the test development process are described as they occur in the test development cycle. First, we comment on the development of test questions. We describe our observations of ETS' process for writing and reviewing test questions and then document results from HumRRO's independent review of the quality of current test questions. Next, we turn to a discussion of the March 2002 CAHSEE administration based on our observation of the administration at a limited number of sites and on results from questionnaires administered to a sample of testing coordinators. Finally, we conclude with analyses of the consistency of the hand scoring of the essays and the accuracy of the overall scores from each part of the CAHSEE.

A significant issue in the administration of the CAHSEE is the degree and consistency with which accommodations are provided to students with disabilities. In some cases, where potential methods of accommodation are judged to alter what is being measured, they are labeled "modifications" and the resulting scores are invalidated. We provide an analysis from the operational data files of the frequency of accommodations and modifications and of passing rates for students receiving accommodations.

Quality of the Test Questions

Observation of Test Development

ETS Item Review. The item development cycle had five major points for review and revision: internal, educator, community, Statewide Pupil Assessment Review panel (SPAR), and field test. Staff from ETS conducted the reviews. A member of the HumRRO staff observed two educator and community reviews.

The purpose of the first educator review was to replenish the item bank for ELA prior to a special January 2002 field test. Several passages had been developed and reviewed under the previous development contract but had too few associated questions (typically only three) for the field test—the ETS goal was six to nine questions per passage. A member of the ETS staff conducted a brief overview session that covered the following characteristics of acceptable questions:

- Match to content standard and construct
- Match to item specifications
- One correct answer
- Plausible distracters
- Appropriate difficulty
- Representative of classroom content

Thirteen current or former ELA educators first reviewed passages to determine whether they were suitable for field-testing. Three passages were deleted. For the remaining passages, educators worked in two groups to review the questions. Most of the new questions were accepted, though reviewers expressed reservations about the literary quality of some of the passages and the appropriateness of a few of the questions, especially those that asked why the author wrote the passage.

The second educator review was one in a series of four, prior to the March and May 2002 field tests. The review addressed both math and ELA items; the HumRRO observer attended the ELA review. The main variation in the method from the earlier review was increased attention to how well each question matched the intended content standard. Overall, all passages were accepted and reviewers were generally positive about the quality of questions.

After each educator review, ETS conducted a community review, which focused on fairness. The key issue was ensuring equal opportunity for students to demonstrate what they know (e.g., avoid regional language and terms). Guidelines included:

- Avoid reinforcing stereotypes (e.g., doctors as men, nurses as women). Check assumptions of knowledge (e.g., feel of snow).
- Avoid sensitive topics that would be acceptable if teacher were present (e.g., war).

Most of the reviewers in the first session were educators. One of the few changes made to eliminate bias was a recommendation to revise a passage that advocated a government action so students would be more familiar with the level of government (community replaced county). Note that changes were only considered for non-copyrighted passages.

Reviewers in the second session were an American Indian lawyer, the president of the California Black School Board, a Hispanic school board member, an advocate for developmentally disabled citizens, and a PTA member/business owner. Panelists completed an orientation exercise in which they reviewed ELA and math questions that exhibited statistical bias. For most of the time, they worked independently. Discussion identified three changes to ELA questions to eliminate possible sensitivity and fairness concerns.

Cognitive Labs

ETS conducted cognitive labs at four high schools after the May 2002 CAHSEE administration. At each site, four members of the ETS staff interviewed students who completed a small set of test questions. Questions were grouped into three sets each for mathematics and ELA, with 13 questions in each of the math sets and about 9 multiple-choice questions and one constructed response (essay) question in each of the ELA sets. Questions represented a range of difficulty. Students were identified by their school to be as representative as possible of the overall population. HumRRO observed two math and two ELA interviews.

Issues addressed were similar for math and ELA:

- What is the question asking?
- What is the first thing you thought of when you read the question? What were you thinking about as you answered the question?
- Multiple choice:
  - Why did you choose that option instead of others?
  - Were there any other options you thought about choosing? Why did you think about choosing that option?
  - Did you use the graphic (if math item had one) or passage (ELA)
- ELA constructed response:
  - What is the prompt asking?
  - How much would you write?
  - What would you write about?
  - How would you use the passage (such as quotations)?
- How difficult was the question for you to answer (easy, medium, or hard)?
- What made it (easy, medium, or hard)?
- Were there any words you did not know?
- Was there anything in the question that you thought was confusing or that bothered you in some way?
- Do you have any ideas as to how the question could be improved?

According to the test developers, the main impact of the cognitive lab is summative information about individual test questions, which can be used during reviews after the field test. The results also give general guidance to item writers on vocabulary and style conventions (for example, all capitals for modifiers such as BEST). The results will also affect how the difficulty of future forms is set. For example, the labs have shown that the difficulty of math items is largely a function of the number of steps in a problem.

HumRRO's Independent Item Review Workshops

Item review workshops were conducted by HumRRO as part of the independent evaluation on May 22 in Sacramento and May 23 in Ontario, California. District and school educators in the ELA and mathematics curriculum were recruited to participate in a one-day workshop to review and rate test questions for alignment to standards. Similar workshops were conducted in May 2000 at the onset of test development and the quality of the questions reviewed at that time was judged to be good. The purpose of the new reviews was to compare the quality of recent CAHSEE test questions, some developed by AIR and some by ETS, to the quality of the earlier questions included in the May 2000 review.

Panelists. A total of 43 teachers and curriculum specialists participated in HumRRO's May 2002 Independent Item Reviews. The panelists were nominated by our points of contact (usually an assistant superintendent) in districts participating in our annual surveys or other data collections. Table 2.1 shows the position of the panelists for each subject. Table 2.2 shows the number of participants by subject, site, and years of experience.

TABLE 2.1 Current Positions of Item Review Panelists

Current Position	ELA Panelists	Mathematics Panelists
Teacher	13	12
Department Chair	2	3
Curriculum Coordinator	1	2
Assistant Principal	2	2
Unknown	2	3
TOTAL	20	22

TABLE 2.2 Years of Experience of Item Review Panelists

		Years of Teaching Experience			
Subject	Site	Less than 5	5–9	10–19	20 or more
ELA	Sacramento	4	2	4	1
	Ontario	2	4	2	1
Math	Sacramento	2	4	5	2
	Ontario	4	1	4	1
TOTAL		12	11	15	5

Selection of Test Questions. We selected (from two different sources) a sample of test questions that matched the content of an operational test form. First we selected questions used in the March 2002 test form. These were older questions originally developed by the prior test development contractor (AIR). We then selected additional newer questions that were field tested (administered to a sample of students, but not used in operational scoring) with the May 2002 administration. These questions were developed by ETS. Table 2.3 shows the number of older and newer questions selected from each content strand and also the number of questions in each operational test form.

The May field test for mathematics did not include many questions from some content strands; a sufficient number of questions from these strands had been tried out previously.

Consequently, more questions from the March operational form were selected to maintain complete coverage of each content standard. For ELA, the May 2002 field test included more questions for each reading passage than are used operationally. This was done so that if some questions had poor statistical properties, it would not be necessary to discard the whole passage. Since we did not know which questions would be kept for each passage, we reviewed all of them. Thus, except for "Writing Conventions," which was assessed with separate multiple-choice questions, we concluded with more questions than are used in an operational form. We also wanted to review extra essay questions because of the importance given to each of these questions in determining the overall score.

TABLE 2.3 Questions Selected for Independent Review

Strand	Older Questions	Newer Questions	Total	Number in Operational Forms
English-language arts				
Reading Vocabulary	6	8	14	10
Reading for Information	11	18	29	24
Response to Literature	12	26	38	24
Writing Conventions	7	6	13	13
Writing Strategies	4	10	14	11
Writing Applications	2	5	7	2 essays
TOTAL	42	73	115	82 MC* + 2 essays
Mathematics				
Grade 6 Prob & Stat	2	4	6	6
Grade 7 Prob & Stat	4	2	6	6
Number Sense	15	1	16	14
Measurement & Geometry	11	8	19	17
Algebra and Functions	16	5	21	17
Algebra 1	8	4	12	12
Math Reasoning**	7*	1*	8*	8
TOTAL	56	24	80	80

* *MC = multiple choice

** Mathematical Reasoning questions are also classified into one of the other content strands. Counts of the questions reviewed show the alternative breakouts. These questions are not counted twice in the totals.

Rating Procedures. Each workshop began with an overview of the CAHSEE, the independent evaluation, and the goal of this independent review of test questions. Panelists were then split into two rooms, one for the ELA panelists and the other for the mathematics panelists. The rating procedures were described and each group rated questions in a training booklet that consisted of released questions. Panelists then discussed their ratings. Panelists with divergent ratings were given an opportunity to describe their rationale for the ratings that they assigned. The discussion was sufficient to ensure that each panelist understood the rating scale, but there was no attempt to force consensus.

For each test question, the booklet included the full text of a content standard followed by the question, the answer choices and a key indicating the correct answer. The rating scale

printed at the bottom of the page is shown in Figure 2.1. The scale described three levels plus two "in-between" categories. Two types of problems were described in the training:

1. Questions that students may be able to answer correctly without having mastered the target standard
2. Questions that require extraneous knowledge or skill so that students who have mastered the target standard may still not be able to answer correctly

Panelists were asked to write specific comments about individual questions in the test booklets. These comments were subsequently reviewed and summarized by project staff.

How well does this question measure the target standard? (Circle one number.)

1	2	3	4	5
Not at all – no relation between mastery and student answers		**Weakly** – students with incomplete mastery may know the answer; those with mastery may answer incorrectly		**Strongly** – students who have mastered the standard will answer correctly; those who have not will not know the correct answer.

Figure 2.1. Rating scale for the independent review of test questions.

For about three-quarters of the questions, one content standard (the actual target standard) was listed. For about one fourth of the questions, we asked panelists to rate how well the question measured each of four different content standards. There were two reasons for this alternative design. First, we wanted to know whether panelists could pick out the correct standard if it were not provided, but thought that it would take too long for them to go through the entire list of possible standards. We selected one alternative standard from the same content strand as the target standard and two alternate standards from a different strand. The idea was to see if the panelist could pick out the correct target standard from this list. The other reason for this design was that we really did want to know whether questions might measure more than one standard. For mathematics, in fact, some questions are designed to measure a mathematical reasoning standard as well as one of the other content standards. All eight of the mathematical reasoning questions in our sample were assigned to this "multiple standard" condition. Panelists were given the target content standard and an alternative from the same content strand and the target mathematical reasoning standard and an alternative mathematical reasoning standard as the four alternatives.

The mathematics questions were organized into four rating booklets, labeled A through D, with 20 questions each. Booklets A, B, and C each included one standard per question and Booklet D included four standards for each question. One third of the panelists began with Booklet A, one third with Booklet B, and one third with Booklet C. When panelists completed one of these booklets they were given the next, until they had completed all three of the single-standard booklets. After lunch, panelists were introduced to booklet D and all worked on this booklet at the same time.

Six rating booklets were used with ELA. The first three, A through C, included reading passages and questions, each with a single standard to be rated. Booklet D contained writing strategy and writing convention questions, each with a single target standard. Booklet E contained both reading and writing questions with four alternative standards for each question. As with mathematics, we included two standards from the target strand and two from an alternative strand. The final booklet, Booklet F, contained essay questions.

Results and Discussion. For each question reviewed, the average rating was assigned to one of four levels. The most notable division was at 3.5 on the rating scale. Questions with average ratings of 4 or 5 were judged to be strong measures of the targeted standard. We further separated questions with average ratings above 4.5 as "very strong" measures of the targeted construct. Below 3.5, we identified questions with average ratings between 2.5 and 3.5 as "weak" measures of the targeted standard and questions with average ratings below 2.5 as not measuring the standard at all.

Table 2.4 shows the distribution of average ratings for the older (March 2002 operational) and newer (May 2002 field test) questions. Overall, 80 to 83 percent of the test questions reviewed were judged to be strong or very strong measures of the targeted content standard. The mathematics questions received somewhat higher ratings than the ELA questions, with 37 percent of the mathematics questions judged to be *very strong* measures of their standard, compared to 27 percent for ELA.

TABLE 2.4 Average Alignment Ratings for Older and Newer Test Questions

Match to Target Content Standard	Percent of Items*		
	Older	Newer	Total
English-language arts			
Very Strong (4.5–5.0)	21	30	27
Strong (3.5–4.4)	60	51	54
Weak (2.5–3.4)	14	18	17
None (1.0–2.4)	5	1	3
Mathematics			
Very Strong (4.5–5.0)	37	38	37
Strong (3.5–4.4)	45	46	45
Weak (2.5–3.4)	11	17	13
None (1.0–2.4)	6	0	5

* Note: Percents may not sum to 100 due to rounding.

As with the questions reviewed in May 2000, the newer questions had not yet been screened on the basis of statistical analyses of results from a field-test administration to a sample of students. The older questions, however, had not only been screened based on field-test results, but had also been selected for operational use and subject to additional review. We are concerned that *any* of these questions were judged to be less than strong measures of the intended content. It is important to note, however, that the panelists in our workshop were not any more expert than the many reviewers engaged by the test developers to comment on the questions; our reviews were intended as an independent check, but not necessarily the "final word" on any particular question.

We looked at the ratings by content area to see the types of standards for which coverage by the test questions was most often problematic. Tables 2.5 and 2.6 show the percentage of questions from each content area judged to be strong or very strong measures of their target standard. Results from the item review workshop held in 2000 are shown for comparison.

TABLE 2.5 Percent of Test Questions Judged to be Strong Measures of Their Targeted Content Standard—English Language Arts

Content Area (Strand)	Percent Judged to be Strong Measures (> 3.5)	
	2000 Item Reviews	2002 Item Reviews
Reading Vocabulary		
Word Analysis, Fluency, and Vocabulary Development (RV)	93	100
Reading Comprehension		
Focus on Informational Materials (RI)	65	76
Literary Response and Analysis (RL)	77	71
Writing		
Writing Strategies (WS)	88	79
Writing Conventions (WC)	97	92
Writing Applications (WA) [Essays]	0*	100
TOTAL	77	81

* Scoring rubrics for the essay questions were not available for the 2000 Workshops and panelists could not judge the effectiveness of the essays without more information on scoring. Extensive information on the essay score scale was available in the 2002 Workshops.

TABLE 2.6 Percentage of Test Questions Judged to be Strong Measures of Their Targeted Content Standard—Mathematics

Content Area (Strand)	Percentage Judged to be Strong Measures (> 3.5)	
	2000 Item Reviews	2002 Item Reviews
Grade 6 Statistics, Data Analysis, and Probability (P6)	98	83
Grade 7 Statistics, Data Analysis, and Probability (P7)	83	67
Number Sense (NS)	95	88
Algebra and Functions (AF)	90	86
Measurement and Geometry (MG)	94	89
Algebra I (A1)	93	92
Mathematical Reasoning (MR)	79	50
TOTAL	91	83

For ELA, the overall alignment ratings were slightly higher than in the May 2000 Workshops, due primarily to more complete information from which to judge the essay questions. (In May 2000, scoring rules had not yet been developed for each essay and, prior to the field test, sample student responses were not available.) For mathematics, the May 2002 ratings were slightly lower than in May 2000, but still somewhat higher than the ratings for the ELA questions.

Table 2.7 shows results from the special booklets designed to see whether panelists could identify the target standard from a list of four alternative standards and whether some questions might, in fact, indicate mastery for more than one standard. The results indicate that most of the questions were judged to be strong or very strong matches to the target standard, and hardly any matched the nontarget standards at all. For ELA, it made little difference whether the target standard was explicitly identified or merely one of four alternatives. There were significant differences for mathematics, but this was primarily due to the fact that all of the mathematical reasoning questions were included in the four-alternative format.

TABLE 2.7 Average Alignment Ratings for the Target and Other Standards

Match to Target Content Standard	Relationship of Standard to the Question*		
	Identified as the Target	Target Standard in List of Four	Other (Nontarget) Standard
English Language Arts			
Very Strong (4.5–5.0)	26%	32%	2%
Strong (3.5–4.4)	56%	45%	7%
Weak (2.5–3.4)	15%	23%	23%
None (1.0–2.4)	3%	0%	65%
(Number of Questions)	(93)	(22)	(66)
Mathematics			
Very Strong (4.5–5.0)	43%	23%	0%
Strong (3.5–4.4)	48%	38%	6%
Weak (2.5–3.4)	8%	23%	6%
None (1.0–2.4)	0%	15%	89%
(Number of Questions)	(60)	(26)	(54)

* Note: Percents may not add to 100 due to rounding.

We conducted further analyses of each of the questions that were not judged to be a strong measure of their targeted content standard. We examined written comments for each of the 22 ELA questions and each of the 13 mathematics questions with low ratings (two of which had multiple target standards with low ratings). We also reviewed notes from the group discussions at the end of each workshop for further explanations of the concerns reflected by the written comments.

Table 2.8 summarizes the main comments on each of the ELA questions that had average ratings below 3.5. Five questions received low ratings because the panelists said that the type of passage did not match the type of standard. The greatest concern, as reflected by average ratings of 2.3, was with the use of information passages to measure response to literature standards. Panelists judged that five other questions assessed a standard different from the target standard. Four of the questions were judged to have flaws, such as requiring information not in the text or being able to be answered without reading the text. The remaining eight questions with low ratings appeared to result from differences between the published test blueprints and the specifications given to item writers. The item specifications allowed coverage of lower-level enabling skills related to target standards in the blueprint. The most common examples of error were questions based on a single passage that were tied

to standards calling for extending or comparing works. The item specifications also allowed genres, such as poetry, that were not explicitly included in the content standard for the grades covered by the CAHSEE blueprints.

Table 2.9 shows the main comments given on mathematics questions that received low alignment ratings. Two questions were given very low ratings because the difficulty of the text may mislead students who have, in fact, mastered the targeted skill. Six more questions were given low ratings because a different content standard was judged to be a better match. Five questions were judged to contain flaws, besides problems with text difficulty. These include questions judged to be so easy, that a correct answer did not indicate mastery of the content standard, and questions that had extraneous or distracting information. The final two questions were targeted to standards involving multiple steps, but were judged to require only one step. These questions may have been akin to the foundational questions described above for ELA.

TABLE 2.8 Reasons for Low Content Alignment Ratings – English Language Arts

Primary Reason for Low Ratings	Number of Questions	Average Rating	Number of Comments
Information passage used for a response to literature content standard	3	2.3	24
Literary passage used for a response to information content standard	2	2.6	7
Better match to a different standard in the same content strand	2	2.8	10
Better match to a standard in a different content strand	3	3.1	10
Item Flaw – requires extraneous information or does not require reading the passage	4	3.2	14
Question measures enabling skill, not the standard as stated in the test blueprint	4	3.2	14
Question encompasses a genre not included in the test blueprint	4	2.8	13

TABLE 2.9 Reasons for Low Content Alignment Ratings – Mathematics

Primary Reason for Low Ratings	Number of Questions	Average Rating	Number of Comments
Text difficulty obscures measurement of the target content standard	2	2.5	9
Better match to a different standard	6	2.7	13
Item Flaw – requires extraneous information or does not require reading the passage	5	3.3	31
Question does not require multiple steps as indicated in the content standard	2	3.1	12

Summary. The results indicated generally good alignment of the test questions to their target standards. Results were similar to alignment ratings collected in May 2000 on the first batch of test questions. The response to literature strand for ELA and the mathematical reasoning strand for mathematics had the lowest proportion of closely aligned questions. Panelists judged the test questions to be clearly aligned to their targeted content standard and not aligned to other, non-target standards. Analyses of reasons given for low ratings of alignment to targeted content standards indicated a number of different reasons for both ELA and mathematics questions. Many of the ELA questions may have been given lower ratings because they matched standards from earlier grade levels that were "foundational" for the indicated target standard. One concern about some of the mathematics questions is that text complexity levels were unnecessarily high. A copy of the questions reviewed, the alignment ratings for each question, and a compilation of the panelists' comments are being given to CDE for follow-up action.

Test Administration

Observation of the March 2002 Administration

For most schools, the May 2004 administration of CAHSEE was their second experience with the test procedures. The schools' increased experience plus CDE improvements in the testing procedures (and possibly also the reduced number of students to be tested) resulted in a test administration that generally worked quite well. In this section we describe sources of information about the administration, the quality of preparation for test coordinators, the impact of logistics on the quality of testing, and the status of accommodations and modifications.

Sources of Information

HumRRO collected information on administration of the CAHSEE from two primary sources: observation at six schools as they administered the CAHSEE and surveys from a sample of school site test coordinators. We also analyzed information on accommodations from the operational March 2002 CAHSEE data files.

Characteristics of the test sessions observed are shown in Table 2.10. The small numbers tested are consistent with the scope of students targeted. The March dates were for 10^{th} graders who had not passed the test for the subject in 2001 in these schools. The May dates were primarily for schools that could not administer the CAHSEE in March, mainly year-round schools, but some schools used the date for make-up tests—for 10^{th} graders who had not passed the test for the subject in 2001 and had not taken it in March. School types D, E, and F were in the make-up category. During an observation, a member of HumRRO's staff interviewed the test coordinator and watched students take the test—attending to the pace of progress, test security, and level of distraction.

TABLE 2.10 Characteristics of Schools Observed

School	Subject	School Type	Approx. Number of Students Tested.	Testing Environment	Accommodations Observed*
A	ELA 1, March	Small town	100	Classrooms	None
B	ELA 2, March	Small town	30	Library	Separate Room
C	Math, March	Rural	200	Library/Cafeteria	None
D	ELA 1, May	Small city	40	Classrooms	None
E	ELA 2, May	Alternative	2	Classrooms	None
F	Math, May	Small city	60	Classrooms	None

* We did not have access to IEPs or Section 504 Plans and had no basis for knowing whether any students requested accommodations. Given the small numbers of students tested, it is not surprising that few students needed accommodations.

The survey of teachers and principals in the longitudinal sample of schools included a survey of site coordinators. The site-coordinator survey asked for feedback on training and guidance, students tested, and the general approach to conducting the test. The point of reference for the survey was the March test. We received responses from 42 site coordinators in 17 districts. The most frequent respondent positions were Test Coordinator (20) and Assistant Principal (18).

Preparation

Site coordinators were prepared for the CAHSEE through three sources: Test Administration Training workshops, guidance documents, and district workshops.

ETS conducted Test Administration Training at five workshops. The workshop that HumRRO observed was attended by 250 people. The trainers emphasized four points:

1. CAHSEE requires a lot of work but is very important for students.
2. Test administration is a critical component of test validity.
3. It is not possible to anticipate every situation, so coordinators and administrators must sometimes exercise professional judgment.
4. The test environment should give every student the opportunity to succeed.

The trainers also handed out manuals covering administration and discussed requirements for receiving and transmitting test materials.

In our survey sample, 13 of the 42 coordinators who responded had attended one of the workshops and 2 had watched the video of a workshop. No one listed any problems with the workshop; five participants, including both who had seen the video, cited it as "most helpful."

Almost every coordinator had read the *Directions for Administration,* the *School Coordinator's Manual,* or both. Their opinions were generally positive—12 cited them as the

most helpful sources of information, especially the scripts for the test administrator. One comment did ask for a more positive introduction to precede the warning against cheating. Other suggestions were to clarify how to treat students who miss one day of ELA and provide guidance on what to do with students who finish early or need more time.

District workshops were the most frequently cited sources of helpful information. Twenty-six coordinators said they had attended such workshops and 14 considered them the most useful source of information, largely because of the chance to ask questions and request follow-up guidance from the district. In contrast, one of the observed sites received little help from the district; for example, they did not know whether they would test in March or May until they got the booklets the week before the March test date.

Logistics

The observations and surveys provided information on five aspects of logistics:

1. type of test facility
2. security
3. the impact of the revised schedule
4. the effect of the consolidated answer document
5. problems encountered

The question about test facility was whether schools tested in a large room, such as a library, cafeteria, or gymnasium, or in classrooms. During March, schools were about evenly divided: Of the 35 coordinators who answered the question, 17 tested in classrooms and 18 tested in large rooms (3 of them off-site). Part of the popularity of the large-room approach is that it is easier to find enough people to conduct the test. The basic requirement is that each session must have a test administrator (with a credential) for the first 25 students plus a proctor (school employee) for each additional 25. An additional reason to favor large rooms in 2002 was that, because about half of the 10th graders had passed the test and were therefore using the regular classrooms, it was hard to find available classrooms or teachers. About a third of the schools that have firm plans for 2003 will test in large rooms.

The extensive use of large rooms is notable because the model configuration for the guidance documents is the classroom. Survey respondents asked for guidance on distance between students and how to treat students who finish early, factors that are especially important in a large-room setting. ("We kept all students together for the testing time but, in the large setting, were just unable to keep 500+ 15-year-olds quiet for that long.") Observation suggests that the schools that use libraries could benefit from guidance on the orientation of tables and the number of people at each table. Many of the issues with large-group testing relate to the balance between providing comfortable test conditions and maintaining security. We added an item on security-related concerns to the questionnaire. Most who responded (22 of 28) thought that they had no security issues. The dominant theme among those who had concerns was that the security procedures were tedious.

ETS, through a subcontract with a security firm, conducted audits of security before, during, and after test administration. The HumRRO observer went through the monitor orientation and completed an observer's test-day report for each site observed. The main concern during the audits was the security of test materials, which was high at all observed sites. In addition, the observer's evaluation guide included several checkpoints on good testing practices that were not included in the guidance documents. These practices included assigning seats randomly, providing adequate workspace, distributing and collecting forms individually in serial order, as well as previously mentioned large-group issues, such as distance between students. Guidance on such practices would enhance the *Directions for Administration.*

The major change in the schedule was to conduct the two ELA sessions on separate days. Most (23) of the coordinators with an opinion on the change favored it, mainly because it was less stressful on students than the "deadly 1-day format." One conclusion from the observations was that the 2-day approach made it much easier to provide additional time. The most common complaint among the 12 coordinators who preferred the 1-day approach was that they lost students between sessions.

Even though there was no effort to solicit comments on the schedule for the math test, one coordinator and one principal recommended testing math over 2 days. Two coordinators recommended testing math before ELA.

Despite repeated explanations from the CDE about restrictions on Saturday testing, it is still a popular requested alternative. About 60 participants (an estimated 25 percent) in the Test Administration Training session HumRRO observed expressed a desire to test on Saturday. One of the principals in the observed schools also made that request, preferring it even if it were voluntary and took 3 weeks.

Another change in logistics from 2001 was the consolidated answer document, which has one section for background information and space for answers to each test. Again the reaction of coordinators was highly positive. Thirty (of 38) coordinators thought the consolidated answer document was more efficient. Dissenters cited the complexity of giving the math test when many had not taken the ELA test, the possibility of students' marking in the wrong sections, and difficulties of reconfiguring groups for testing.

The survey included a question about problems that were not covered by guidance documents. The most frequent comment (seven coordinators) was about a mismatch between the number of digits of test booklet number and the answer document. CDE had learned of the problem early and sent an e-mail that morning notifying coordinators of the problem. Some test administrators, including one at observed school A, did not get the information before the test session started. At the observed site, the confusion was momentary and did not affect the test conditions.

Accommodations and Modifications

Accommodations include changes to test presentation, response, or scheduling to provide a more appropriate assessment of students with disabilities. Modifications are changes that also change what is being measured and so invalidate the resulting test scores. According to

CDE regulations, the decision to grant accommodations or allow modifications must be based on the student's Individual Education Program (IEP) or Section 504 Plan. Students whose plans require test modifications cannot pass the exam directly, but may apply for a waiver if their test scores and other evidence suggest that they have mastered the required skills.

The SBE passed regulations regarding accommodations and modifications and the CDE began disseminating information about the regulations through workshops for special education coordinators and three regional workshops for special education coordinators and test coordinators. After the workshops, the CDE distributed an extensive *CAHSEE Accommodation Training Manual* through district and county superintendents to each school site. Two specific modifications were identified as invalidating score results in the CAHSEE regulations: calculator use on the math test and an audio or oral presentation of the ELA test. Districts were asked to submit questions about additional types of accommodations required in students' IEP or Section 504 Plans to the CDE. During the subsequent review process, authorized by the California Code of Regulation (Section 1218), other changes were identified as modification for the Constructed Response items on th eELA portion: transcriber, mechanical or electronic transcribing devices, and spell-check devices/software.

Before the March test date, the Federal District Court required the CDE to distribute a notice to districts to be sent to all parents and guardians of students with an IEP or 504 Plan. The notice announced the student's right to accommodations and modifications that are permitted in any instruction (not just standardized testing) and specified that no additional IEP or 504 Plan team meeting was necessary.

While not frequent in the schools observed, accommodations and modifications were more common in the surveyed schools. The most frequent accommodations were setting, such as separate room (24 schools) and timing/scheduling, such as more frequent breaks (23 schools). Presentation accommodations were three large print versions, one Braille transcription, and six audio or oral presentations of the math test. Response accommodations were two verbal, written, or signed responses and one assistive device that does not alter content. In the modification category, ten schools allowed some students to use calculators for math, five allowed audio or oral presentation for ELA, and one allowed assistive devices judged to alter the content of the measures (spell-check of writing and word processor).

Survey responses suggest that a minority of schools are still unsure how to deal with special education students with regard to the CAHSEE. Nine schools estimated that they tested fewer than half of their eligible special education students; three of those nine tested none of them. Some of those schools may have focused on special education students during May; that was the plan for at least one district in our sample. Another possibility is that schools may be confused by other state programs that exempt some special education students from testing.

In addition to the survey responses, we examined data files from ETS on all of the students tested in March 2002. Students often received more than one accommodation. When this happened, we identified one of their accommodations or modifications as "primary" so as to avoid double counting. We also wanted to avoid confusion about cases

where students receiving an accommodation also receive an invalidating modification. Tables 2.11 and 2.12 show the frequency of different "primary" modifications and accommodations and also the most common variants of accommodation for each primary accommodation.

Where students received more than one modification or accommodation, the rules for identifying a primary accommodation were as follows:

1. Modifications took precedence over any accommodation.
2. A named modification or accommodation took precedence over any of the "other" categories (except that "Other modifications" took precedence over any accommodation, following Rule 1).
3. Within named accommodations (or "others"), presentation accommodations were taken first, then response accommodations, and then scheduling accommodations. Scheduling accommodations were the most common and were judged to have the least effect on scores.

TABLE 2.11 Frequency of Accommodations and Modifications—ELA

Primary Modification or Accommodation	Frequency		Most Common Other Modifications or Accommodations			
	Number of Students	Percent	Most Frequent	Number	Next Most Frequent	Number
Modifications	**918**	**0.54**				
Audio Presentation	675	0.40	Directions Read	360	Additional Time	358
Other Modification	243	0.14	Audio Presentation	243	Directions Read	80
Accommodations	**7,001**	**4.14**				
Presentation	*2,553*	*1.51*				
Braille	22	0.01	Scribe	15	Directions Read	13
Large Print	67	0.04	Additional Time*	28	Scribe	20
Directions Read	2,353	1.39	Additional Time*	1335	Additional Breaks	839
Other	111	0.07	Other Response	14		
Response	*197*	*0.12*				
Scribe	45	0.03	Additional Time*	16	Answer in Book	10
Answer in Book	120	0.07	Additional Time*	52	Additional Breaks	43
Other	32	0.02	Other Scheduling	1		
Scheduling	*4,251*	*2.51*				
Additional Time*	2,074	1.23	Other Scheduling	125	Other Presentation	48
Additional Breaks	1,895	1.12	Additional Time*	1468	Other Scheduling	169
Other	282	0.17				
None	**161,231**	**95.32**				
TOTAL	**169,150**	**100.00**				

*Note: Additional time is allowed for all students.

For ELA, fewer than five percent of the students tested received any accommodation or modification. Only about one-half of a percent took the exam with a modification. The most frequent accommodations were having the directions read, additional time, and additional breaks. Additional time and additional breaks were usually offered together.

TABLE 2.12 Frequency of Accommodations and Modifications—Mathematics

	Frequency		Most Common Other Modifications or Accommodations			
Primary Modification or Accommodation	Number of Students	Percent	Most Frequent	Number	Next Most Frequent	Number
Modifications	**4,333**	**1.85**				
Calculator	4,277	1.83	Additional Time*	1256	Directions Read	990
Other Modification	56	0.02	Other Presentation	21	Additional Time	19
Accommodations	**5,154**	**2.20**				
Presentation	*1,725*	*0.74*				
Braille	22	0.01	Scribe	12	Answer in Book	8
Large Print	65	0.03	Additional Breaks	19	Scribe	17
Audio Presentation	104	0.04	Additional Time*	22	Additional Breaks	16
Directions Read	1,415	0.60	Additional Time*	773	Additional Breaks	583
Other	119	0.05	Other Response	16	Other Scheduling	10
Response	*137*	*0.06*				
Scribe	25	0.01	Additional Time*	9	Additional Breaks	5
Answer in Book	108	0.05	Additional Time*	29	Additional Breaks	21
Other	4	0.00	Other Scheduling	1		
Scheduling	*3,292*	*1.41*				
Additional Time*	1,450	0.62	Other Scheduling	81	Other Presentation	35
Additional Breaks	1,513	0.65	Additional Time*	1132	Other Scheduling	116
Other	329	0.14				
None	**224,641**	**95.32**				
TOTAL	**234,128**	**100.00**				

*Note: Additional time is allowed for all students.

We were surprised by the large number of students who took the math test with an invalidating modification. Nearly two percent were allowed to use calculators. As with ELA, having the directions read, additional time, and additional breaks were the most common accommodations. As with ELA, fewer than five percent of the students taking the math test received any accommodation or modification.

We also examined the relationship between a student's primary disability and the type of accommodation he or she received. Tables 2.13 and 2.14 show the number of students with each type of disability for each type of accommodation.

TABLE 2.13 Number of Students With Modifications or Accommodations* by Primary Disability—ELA

	Primary Disability Code*											
Primary Modification or Accommodation	090 L.D.	060 E.D.	010 M.R.	120 Aut.	130 B.I.	020 030 Deaf	050 100 Vis.	040 S/L Imp	070 Orth.	080 110 O/M	Unk	None
Modifications												
Audio Presentation	497	8	31	3	0	3	5	20	7	15	4	82
Other Modification	161	14	4	2	2	1	3	8	9	11	1	27
Accommodations												
Presentation												
Braille	1	0	0	0	0	0	13	0	0	0	0	8
Large Print	14	0	5	0	1	0	24	0	5	3	1	14
Directions Read	1,524	81	62	10	11	107	8	88	19	66	29	348
Other	76	10	2	3	0	1	0	4	0	1	0	14
Response												
Scribe	11	4	0	1	0	0	0	0	11	5	1	12
Answer in Book	51	22	3	1	0	0	0	1	2	6	6	28
Other	10	0	0	0	0	0	0	0	1	3	0	18
Scheduling												
Additional Time**	1,373	100	18	8	3	9	0	36	9	65	12	441
Additional Breaks	1,244	214	31	9	3	11	0	36	5	51	25	266
Other	194	18	6	0	0	2	0	11	1	7	4	39
None	13,066	962	328	58	35	135	59	874	107	577	285	144,745
Not Tested	3,723	260	57	15	4	49	34	232	37	203	79	78,430
TOTAL	**21,945**	**1,693**	**547**	**110**	**59**	**318**	**146**	**1,310**	**213**	**1,013**	**447**	**224,472**

*Key for disability codes:

090 L.D.—Specific Learning Disability
060 E.D.—Emotional Disturbance
010 M.R.—Mental Retardation
120 Aut.—Autism
130 B.I.—Traumatic Brain Injury
020, 030 Deaf—Hard of Hearing, Deaf
050, 100 Vis —Visual Impairment, Deaf-Blindness
040 S/L Imp.—Speech or Language Impairment
070 Orth.—Orthopedic Impairment
080, 110 O/M—Other Health Impairment, Multiple Disabilities
Unk.—Disability indicated, but no code given
None—None

**Note: Additional time is allowed for all students.

TABLE 2.14 Number of Students With Modifications or Accommodations* by Primary Disability—Math

Primary Modification or Accommodation	Primary Disability Code* 090 L.D.	060 E.D.	010 M.R.	120 Aut.	130 B.I.	020 030 Deaf	050 100 Vis.	040 S/L Imp	070 Orth.	080 110 O/M	Unk	None
Modifications												
Calculators	2,796	99	84	14	6	35	13	148	38	110	56	878
Other Modification	25	10	2	1	0	0	1	4	1	2	0	10
Accommodations												
Presentation												
Braille	1	0	0	0	0	0	13	0	0	1	0	7
Large Print	17	0	4	0	1	0	25	0	3	3	1	11
Audio Presentation	72	4	6	1	0	0	1	2	1	3	0	14
Directions Read	876	56	58	4	6	90	9	51	16	44	16	189
Other	76	12	2	3	0	0	0	4	0	2	0	20
Response												
Scribe	7	2	0	0	0	0	0	0	7	2	0	7
Answer in Book	39	20	2	1	0	0	1	1	3	3	5	33
Other	1	1	0	0	0	0	0	0	0	0	0	2
Scheduling												
Additional Time**	958	93	18	5	3	5	2	15	6	54	4	287
Additional Breaks	976	191	18	7	2	11	1	30	5	39	18	215
Other	218	17	8	0	0	5	0	15	1	6	2	57
None	14,426	1,068	313	68	35	153	75	977	121	675	304	206,426
Not Tested	1,457	120	32	6	6	19	5	63	11	69	41	16,316
TOTAL	**21,945**	**1,693**	**547**	**110**	**59**	**318**	**146**	**1,310**	**213**	**1,013**	**447**	**224,472**

* See Table 2.13 for Key for disability codes

**Note: Additional time is allowed for all students.

By far, the most common primary disability category was specific learning disabilities (LD). Two-thirds of these students completed the math exam with no accommodations or modifications, but nearly 13 percent of them had a calculator modification. Further investigation may be warranted for the modest number of cases where modifications or accommodations were allowed but no disability was indicated.

For the ELA exam, specific learning disabilities were also indicated for the majority of the students taking the test with a modification. For both tests, it appeared that directions were read for a modest number of students with no indicated disability.

A key question was how many of the students who received an accommodation passed the corresponding part of the CAHSEE and also how many students who took one or both parts with a modification received a score of 350 or more. Table 2.15 shows the percent of students scoring 350 or above (passing) for each type of modification or accommodation.

TABLE 2.15 Percent Scoring 350 or Above by Primary Modification or Accommodation*

Primary Modification or Accommodation	Students Scoring 350–450 ELA N	ELA %	Mathematics N	Mathematics %
Modifications		**23.0**		**9.7**
Audio Presentation	675	20.7	N/A	N/A
Calculators	N/A	N/A	4,277	9.7
Other Modification	243	29.2	56	7.1
Accommodations		**22.2**		**8.1**
Presentation		*18.1*		*7.2*
Braille	22	45.5	22	13.6
Large Print	67	40.3	65	20.0
Audio Presentation	N/A	N/A	104	7.7
Directions Read	2,353	16.8	1,415	5.9
Other	111	27.0	119	14.3
Response		*37.1*		*10.2*
Scribe	45	60.0	25	24.0
Answer in Book	120	21.7	108	7.4
Other	32	62.5	4	0.0
Scheduling		*23.9*		*8.4*
Additional Time**	2,074	24.4	1,450	8.0
Additional Breaks	1,895	24.4	1,513	8.3
Other	282	26.2	329	10.9
None	**161,447**	**55.9**	**224,641**	**33.5**
TOTAL	**169,150**	**54.4**	**234,128**	**32.5**

* Scores for students receiving a modification were invalidated and did not count as passing.
**Note: Additional time is allowed for all students.

Passing rates for most students receiving an accommodation were significantly lower than the overall rates. The one exception was for the relatively small number of students who were allowed a scribe. A modest number (about 23 percent) of students receiving a modification on the ELA test would have passed. For math, fewer than 10 percent of the students who used calculators would have passed if their scores had not been invalidated.

Scoring

For the 2001 administration, essays were scored using the approach used with most large-scale assessments. Scorers were assembled at one or two fixed sites and organized into several groups, or "tables." After initial training and a calibration exercise, paper copies of student essays were distributed and two scorers were assigned to each essay. Table leaders answered questions, periodically checked the accuracy of each scorer through "check sets,"

and attempted to resolve situations where the primary scorers disagreed over the scorabilty of a response or assigned scores that differed by more than one score point. In a few situations in which the table leader did not agree exactly with either of the primary scorers, a fourth, even more expert scorer was brought in to resolve the issues and assign a final score.

For the 2002 administration, ETS proposed and used a notably different system for scoring the essays. Scoring leaders were identified and provided with face-to-face training. Other scorers were recruited and trained "online," usually working from home. Scorers viewed the students' responses on their computer screens rather than in paper form. Scoring leaders were available online and by phone to answer questions. Software was designed to monitor scoring accuracy and provide scoring leaders with immediate notification of emerging problems.

The new ETS approach had significant cost and possibly also time advantages. It also ensured that the two readers for each essay were totally independent. The key question is whether the online training and monitoring used in 2002 was as effective as the in-person training and monitoring used in 2001.

Training Scoring Leaders for ELA Constructed-Response Items

The scoring leaders were key to the online scoring approach for the two constructed response items. Each leader coordinated at least eight readers, contacting each of them at least once a day. The purposes of the session were to train 30 potential leaders on how to score the essays and to model the appropriate mentoring relationship with the readers. The orientation emphasized the relation between the scoring guide and ELA standards, with particular emphasis on writing conventions such as spelling, which received more emphasis in 2002 than in 2001.

Training focused on the two prompts in the March 2002 administration: a stand-alone prompt on teen dress and a response to a literary/expository passage. For each prompt, trainees reviewed four benchmark papers that represent midpoints in each of four levels of the scoring guide. After discussing the benchmarks, they read range-finding papers grouped by upper half and lower half and discussed them. During discussions, a member of the ETS staff polled trainees for their ratings and then explained the "book" rating, trying to maintain the link with the scoring guide.

After about 30 passages, trainees worked in a computer lab to practice working with the online scoring system. Practice included gaining access to topic support written by test developers.

Scoring Consistency

Table 2.16 gives information on scoring consistency for the March 2002 administration in comparison to the March and May 2001 administrations. Scoring problems (disagreement over scorability or score differences of more than one point) were still very infrequent, but did occur more often than last year (about 3 percent this year compared to about 1 percent last year).

Note that differences in scoring consistency may not accurately indicate differences in the effectiveness of the two approaches. The March 2002 examinees were much more homogeneous than in 2001 because high scoring students from 2001 did not have to retake the exam and low-scoring students from 2001 had an additional year of schooling. Thus, it was likely that there were significantly fewer essays that were clearly very good or clearly inadequate, leaving more opportunities for scoring disagreements. Table 2.17 shows the score frequencies for the two essays from each administration to date.

TABLE 2.16 Level of Scorer Agreement for the Two Essays

		Percent Perfect Agreement		Percent Adjacent	Percent Discrepant	
Form	N	Scorable	Not Scorable	Scorable	Scorable Difference > 1	Difference in Scorability
March 2001 Essay 1	349,401	67.5	7.0	24.5	0.6	0.3
March 2001 Essay 2	349,401	55.2	9.7	33.0	1.2	0.9
May 2001 Essay 3	19,986	63.3	7.4	27.9	1.0	0.4
May 2001 Essay 4	19,986	54.3	17.6	26.7	0.5	0.8
March 2002 Essay 5	169,159	56.2	6.0	34.5	2.0	1.2
March 2002 Essay 6	169,159	56.9	4.7	35.6	1.9	0.9

TABLE 2.17 Percent of Students at Each Score Level for the Essays in Each CAHSEE Administration

	March, 2001		May, 2001		March, 2002	
Essay Score	Essay 1	Essay 2	Essay 1	Essay 2	Essay 1	Essay 2
0	7.2	10.2	7.6	18.0	6.7	5.2
1	4.9	6.2	4.3	6.4	22.8	16.4
2	26.9	25.1	36.0	22.4	41.4	43.0
3	53.9	42.0	45.5	38.9	22.3	27.8
4	7.1	16.6	6.6	14.2	6.8	7.7
Total Number of Students	349,401	349,401	19,986	19,986	169,159	169,159

Review of Electronic Training for Online Scoring Process

The following observations and suggestions are summarized from reviews of the electronic training process completed by a member of HumRRO's CAHSEE Outside Consultants Panel who brings particular expertise in state assessments and reading and ELA and who has focused on designing scoring guides and training for their use. Students wrote two essays, one stand-alone writing task and the other a response to a passage. Comments are provided here separtely for each type of essay.

Online Scoring of the CAHSEE Writing Task Component. The expectations for readers during operational scoring seem realistic and appropriate regarding hours, break time, and approximate yield per day. However, there was no evidence that the minimums and averages accounted for time to recalibrate, contact with the scoring leader, and score

embedded read-behinds. It would have been helpful to clarify whether these numbers reflected all readings or only readings of live papers for purposes of assigning a score.

Prospective readers are informed that the scoring leader will keep track of their accuracy, use of the full range of score scale, reading rate, reliability, and "professional behavior." It is not clear how, other than by ensuring that readers have the opportunity to be exposed to samples at all score points, one can monitor use of the full range of the score scale, however. Experience with other programs has demonstrated that the number of responses at the lowest score point diminish, and even disappear, over time as instruction and familiarity with the test format improve. Readers still need to be trained to recognize samples that should be considered non-scorable or assigned a score of 1, even though in operational scoring, they may rarely see such samples. Otherwise, readers might shift to a more purely holistic process—assigning the lowest score point possible to the weakest samples. It should be noted, however, that at this time there are still a significant number of papers at these low score points in this initial CAHSEE cohort.

The guidelines given for CAHSEE "crisis papers" are fairly consistent with other hand scoring programs, and readers are given clear information about when to defer to the scoring leader. One by-product of the true randomization that can take place for electronic delivery of responses is that it may become somewhat more difficult to identify such administrative alerts as possible copying and cheating. Again, it would be useful to know if there are any procedures in place to randomly screen samples with administrative groups for this possibility.

Scoring Training—Process and Content. In many very appropriate ways, the electronic training process mirrors that of a high-quality traditional hand scoring training process. Readers are exposed to (a) benchmark (anchor) papers, (b) rangefinder papers, (c) training sets, and (d) certification (qualifying) sets. Readers are given two opportunities to pass a certification test, which is fairly typical of traditional hand scoring projects.

One of the first components of training is supposed to be a review of the Scoring Guide (the rubric). The background material informs readers that decisions are based on such features as content, logic, organization, development, and attention to conventions of English. At no point, however, are readers walked through the rubric, which is a critical first step in traditional training programs. Readers need to examine and understand the parallelism across score points, clearly seeing the "stepping stone" nature of score point descriptors, and get a fix on the key traits or elements of writing across all score points. It also is important that training include clarification of the relative weighting of traits or elements. If elements are to count equally, that needs to be made explicit. If relative strength or weakness in one area can pull a score up or down, that too needs to be made explicit.

The section on "Scoring Notes" was impressive. These often clarified the latitude given to students' interpretation of the topic, sources from which they could draw examples, and use of inaccurate information. To improve the scoring notes, however, readers need to be apprised of any general guidelines that may apply across prompts, in addition to those that are prompt specific. This will avoid the frustration of being informed of an important guideline in an annotation after making a scoring decision.

The benchmarks were usually excellent examples of solid 1, 2, 3, and 4 scores. However, readers have to click on "comments" to learn why that score was assigned and how each benchmark fits the criteria for that score point. Readers may bypass the comments altogether, drawing their own inferences about why each sample was assigned a particular score. It would be useful if comments were linked so that before proceeding to the next score point, the annotation automatically came up.

The number of benchmark and rangefinder papers seemed sparse. Usually, anchor sets comprise three or four samples at a score point, often demonstrating the range from low to high within a score point, or some various routes toward the same end. Alternatively, anchors are solid samples and subsequent training on "split sets" of "fence-sitters" helps to clarify the lines between score points.

There was only one sample per score point for each of the rangefinders. One of the greatest weaknesses of the training protocol is that readers were not guided within/across the range of samples at a given score point. This seems particularly significant given that "use of the full range" is emphasized in the background material. Again, readers must elect to check the comments. Otherwise they can walk away with very idiosyncratic—and perhaps incorrect—ideas about why a given sample received the score it did. Unlike the benchmarks and rangefinders, for which readers had to click on "show comments," readers always saw the annotations for each training set paper (8 per prompt).

One concern is that annotations often departed from the language of the scoring guide. In training for most hand scoring projects, readers are required and expected to stick to the language of the rubric, or in some cases, to that language and a clearly demarcated set of synonyms or terms that fall under key traits. A number of annotations contained what are sometimes called "weasel words"—vague terms used to make qualitative judgments. Annotations also were erratic in the order of ideas or information included, instead of following the order of traits or features identified in the scoring guide.

None of the prompts included in the training identified an audience, although it is one element in the scoring guide and a number of annotations include comments about attention to audience. One of the operational prompts was corrected to include audience, but this should be clearly addressed in all prompts.

Another concern is the review of training samples without returning to the benchmarks for the next prompt. If the intent is for readers to see the commonalties among prompts and to mix training, then the training samples should be truly shuffled rather than being presented in a set of training samples for each prompt. The construction of the training sets, however, was impressive in terms of variation in order and number of score points represented.

The consistency of scores was not always comfortable, and one sample 4 for the first prompt looked very much like a 3 for another prompt. If the test developer's senior scoring staff have not lined papers up across prompts to ensure consistency, this certainly should be done.

Other Concerns. One "truism" of judgment-based scoring is that it is quite possible to give the right score for the wrong reasons. It is not evident that the training includes any steps to ensure that this does not occur. In high quality traditional hand scoring projects, readers are expected to support their judgments with explanations—to ensure that they are applying criteria correctly. This may be done orally or in writing and is often part of the discussion that takes place as part of the "training set" stage of training. One serious omission is the lack of annotations to explain "true scores" for the calibration set papers. It was possible to predict what the "true score" would be, even with defensible reasons for a different score. It would have been useful to have annotation to help understand why the sample got that score.

Key Recommendations for Writing Task Component

- Add to training protocol a guided review of the actual scoring guide
- Clarify the relationship among scoring guide elements and the weight when making scoring decisions
- Front load all general scoring guidelines
- Front-load all prompt-specific guidelines
- Expand number of benchmarks and rangefinder papers—make sure these illustrate both the range and the variety of responses within each score point
- Decide upon electronic "pathways" through components of training—determine how much choice in order of steps should be left to the trainee
- Revise annotations to incorporate language of scoring guide—eliminate "weasel words" and language not clearly aligned to scoring guide traits
- Revise annotations to stick to order of features of scoring guide, as much as possible
- Provide annotations for sample certification sets

Online Scoring of "Response to Passage" Items. A review also was conducted on the online scoring training for the CAHSEE "response to passage" items. Many of the same issues raised above held for the constructed response (CR) items.

It was noted that most of the writing standards clearly align with traditional domains or purposes for writing: narrative (WA36), expository (WA38), and persuasive (WA39). Writing Standard WA40 (10.2.5): business letters is more of a subset of expository writing, and it is not at all clear how scoring of such items would differ, if at all. However, write response to literature (WA37) strongly suggests reading skills in addition to writing skills. The students' "Checklist for Your Writing," which accompanies each prompt, may not reflect fully the reading skills being assessed.

Scoring Training on Prompts. Prompt #1 ("Seining for Minnows")—The feedback on score decisions balanced reading criteria ("grasp of text") and writing criteria ("technical command of English"). As noted above, the language of the scoring guide was not used fully or consistently; instead, most feedback centered on the first, second, and last bullet at each score point.

The distribution of samples across the scale could be improved. The rangefinder for score point 1 was so low it could not be used. Care should be taken that the benchmarks represent solid examples of each score point, while rangefinders help define the lines. The feedback on several benchmarks and rangefinders raised a concern that synthesis was not being rewarded as much as "lift and pluck."

As in the feedback on the stand-alone prompt responses, notes often departed from the language of the scoring guide, more so for the training samples than the benchmarks and rangefinders. Throughout training, the feedback centers on the organizational plan of the response, even though that dimension/criteria does not appear in the scoring guide.

Prompt #2 ("On Screen")—Perhaps because this prompt connects much more with students' experience than Prompt #1, students were much more likely to integrate ideas and experiences that are linked to the text. Therefore, the impact of non-textual support also needs to be clarified, as it was not well addressed in any portion of training. Doing so is important particularly because it reflects current thinking about reading and reading instruction, and the ways readers construct meaning by making text-to-self connections, text-to-text connections, and/or text-to-world connections.

Key Recommendations for "Response to Passage" Items

- Reconsider decision to score passage-based CR only once
- Revise scoring guide to avoid confounding standards in reading and writing
- Revise prompts to make clear key demands, and make sure these are aligned with the scoring guide to be used
- Based on topic notes generated as a result of field testing, consider revising operational prompts so that it is clear to students which components of the response are expected ("be sure to...") and which are optional ("You may choose to...")
- Enhance (supplement) benchmarks and rangefinders to illustrate the scale more fully within and across score points
- Revise feedback (annotations on score decisions) to stick more completely to the language and criteria of the scoring guide (rubric) and to eliminate more personalized language and criteria

Verification of Test Score Equating

HumRRO received a preliminary data file from ETS the second week in April, in time for review prior to final decisions about scoring and scaling. The most salient feature of this file was that the ELA results were flagged as incomplete because scores were not yet available for the essay questions. HumRRO recommended that the ELA scaling and equating analyses be rerun with more complete data before finalizing the equating and the resulting raw-to-scale score conversion tables. ETS agreed with this recommendation and did rerun the scaling and equating with an essentially complete file.

We computed classical item statistics including analyses of the proportion selecting each response option and the biserial correlation (Clemans-Brogden) of selecting that option with the total number correct score. The results verified the keying of the correct option for each question (significant item-total correlation versus negative correlations for each of the incorrect options). We also examined DIF statistics for Hispanics. A few items had marginal statistics that we would have flagged for further review. ETS did review a small number of questions and found no reason to drop or rekey any of the questions. [For the final report, we will insert a table showing the distribution of p-values and biserials for each test.]

We conducted analyses to provide a check on the ETS equating results. Our approach used a subset of the items used in ETS analyses and a different item response theory (IRT) model. ETS included statistics for items from the spring and fall 2000 Field Tests as well as statistics from the operational March and May 2001 administrations, after performing analyses to put Rasch item difficulty estimates from these different sources on the scale of the March 2001 operational form (the one from which the passing standards were established). We were concerned that students participating in the field tests might have been less motivated in comparison to students in the operational 2001 administrations. A difference in motivation might disproportionately affect some types of questions (e.g. hard ones or questions taking longer to answer), thereby skewing the item difficulty estimates.

We limited our analyses to tryout questions from the March and May 2001 operational administrations where student results counted. We further limited our analyses to tryout questions, which were new questions that did not count in determining a student's score. The tryout questions were included in one of 10 different test forms and so only about 10 percent of the students were exposed to them. Each operational item was administered to all students in a given administration. A significant number of students in the March 2002 administration also participated in the 2001 administrations and might have had some memory of questions from that exam. For tryout questions, at least 90 percent of the repeat examinees would not have seen the question in 2001.

We estimated item parameters for the 3-parameter logistic item response model (3PL), which explicitly models the effects of guessing. For the essay questions, we used Muraki's Partial Credit Model (ref.). We had similar parameter estimates for these same items from the analyses of the March or May item tryouts. Stocking-Lord equatings were performed separately for questions from the March and May tryouts. The resulting linear conversions permit the parameter estimates from March 2002 to be transformed to fit the scale of the May or March 2001 parameter estimates. Since all items were scaled relative to an examinee distribution with mean 0 and variance 1, the linear transformation could also be used to estimate the mean and variance of the March 2002 calibration sample relative to the mean and standard deviation of the March or May 2001 calibration samples. Table 2.18 shows summaries of the item parameter estimates from 2001 and 2002.

TABLE 2.18 Distribution of Item Difficulty and Slope Parameter Estimates

Subject	2001 Admin.	No. of items	Base Administration			March 2002		
			Difficulty		Slope	Difficulty		Slope
			Mean	SD	Mean	Mean	SD	Mean
ELA	March	8	-.89	.82	.72	-.09	.68	.90
	May	26	-.19	.67	.83	-.20	.89	.87
Math	March	8	.05	.81	.87	.31	.75	.94
	May	14	.32	1.00	.94	.48	1.06	1.08

The March 2002 calibration indicated that the ELA items were significantly more difficult than the items used in March 2001 and about the same level of difficulty as those used in May 2001 items. In fact, the items did not change between 2001 and 2002, so the results imply that the March 2002 sample performed at about the same level as the May 2001 sample and at a significantly lower level than the March 2001 sample. The larger slope estimates in 2002 imply that the 2002 sample had a smaller variance, as do the smaller standard deviations of the difficulties in comparison to March 2001.

For Math, the item difficulties appeared greater in 2002 than in 2001 in that fewer students passed the common items. This implies that the 2002 sample performed at a lower level than either the March or May 2001 sample. Again, the higher slope estimates in 2002 suggest that the variance of the 2002 sample was somewhat smaller.

Table 2.19 shows estimates for the March 2002 means and standard deviations derived from the calibration of our sample of items. These estimates are compared to our computation of actual values for the 2002 sample using the conversion tables supplied by ETS. (Note that we checked the initial conversion tables for ELA; we were not able to check the revised tables in the time available.) For ELA, estimates derived from the March 2001 and May 2001 field-test items were somewhat divergent. The estimates from the March 2001 comparison were based on many fewer items (8 versus 26) and on samples that were significantly different in their performance. The May estimates suggested that the March 2002 sample was only very slightly higher performing than the May 2001 sample. The combined estimates weighted the two separate results based on the number of items included and yielded estimates similar to the estimates from the May 2001 field-test items. The values for this sample computed with the ETS conversion table were very close to these combined estimates.

TABLE 2.19 Estimated Scale Score Means and Standard Deviations for March 2002

Subject	2001 Admin.	2001 Sample		Linear Conversion		Est. for March 2002	
		Mean	SD	Slope	Intercept	Mean	SD
ELA	March	365.4	38.4	1.002	-.491	346.6	38.4
	May	355.2	39.7	0.824	.093	358.8	32.8
	Combined (Weighted) Estimates					356.0	34.1
	Values Computed from ETS Conversion Tables					356.3	34.6
Math	March	349.1	36.8	0.939	-.203	341.3	34.6
	May	343.0	37.6	0.846	-.013	342.5	31.8
	Combined (Weighted) Estimates					342.1	32.8
	Values Computed from ETS Conversion Tables					340.4	32.3

The results are highly confirmatory. Our very rough estimates agreed with ETS' values with respect to the direction and generally the amount of differences between the March 2002 and March and May 2001 samples. Thus, differences in the items included and the scaling and equating methodologies did not lead to any appreciable differences in the resulting scale scores.

Test Form Accuracy

The CAHSEE is used to make graduation decisions about students. It also provides feedback as to how far above or below the minimum a student is, but this is a less significant use of test score information. In prior reports (e.g., Wise et al., January 2002), we described a procedure for examining the accuracy with which test scores classify students as either passing or not passing. The procedure is based on defining a zone of uncertainty as the score range for which students have more than a 5 percent chance of being classified incorrectly. There will always be some amount of uncertainty for any test of finite lengths. For students whose true achievement level is very slightly below (or above) the minimum passing scores, the probability is nearly .5 that the student will receive a passing (or not passing) score from any single testing session, so the classification error rate will always approach 50 percent right at the cutoff.

The practical question is whether the zone of uncertainty includes students with true achievement levels that are significantly or only trivially below or above the minimum passing level. Table 2.20 shows the "zone of uncertainty" for the March 2002 administration and the estimated probability of passing the exam in a single testing for each true achievement level. Table 2.20 also shows the percentage of all students at each score level and the percentage of all students who may be incorrectly classified.

TABLE 2.20 Estimated Classification Error Rates for the March 2002 Form

True Level of Achievement	Score Range		Percent in Range	Estimated Percent Passing	Total Percent Potentially Incorrectly Classified
	Percent of Total Points	Scale Scores			
English-language arts					
1. Well Below Minimum	00–48	250–334	28.1	0.6	0.2
2. Slightly Below Minimum	49–57	335–349	17.5	23.2	4.1
3. Slightly Above Minimum	58–64	350–365	17.1	76.9	4.1
4. Well Above Minimum	66–100	366–450	37.3	99.4	0.2
Range of Uncertainty	49–64	335–365	34.6		8.2
Outside this Range			65.4		0.4
TOTAL			100.0		**8.5**
Mathematics					
1. Well Below Minimum	00–46	250–335	47.6	1.1	0.5
2. Slightly Below Minimum	48–55	336–349	19.9	24.1	4.8
3. Slightly Above Minimum	55–63	350–362	12.7	77.9	2.8
4. Well Above Minimum	64–100	363–450	19.8	99.5	0.1
Range of Uncertainty	48–63	336–362	32.6		7.5
Outside this Range			67.4		0.6
TOTAL			100.0		**8.2**

Table 2.21 compares the accuracy analysis for the March 2002 test form to the test forms used in March and May 2001. The most noticeable difference is that significantly more students were in the zone of uncertainty in the March 2002 administrations than in prior administrations (33–35 percent compared to 20–24 percent). For ELA, the zone of uncertainty was slightly wider in March 2002 (possibly because of slightly lower scoring consistency on the essays). Most of the difference, for both ELA and mathematics, however, was a reflection of true differences in the populations who took the exam. The March 2002 administration did not include all of the students who passed previously, most of whom were well above the zone of uncertainty. In addition, a number of the repeat examinees (and also the first-time examinees) benefited from an additional year of instruction and so were less likely to be significantly below the zone of uncertainty. The overall classification error rates were potentially higher in March 2002 because more students were near the cutoff. The percent of students who were outside the zone of uncertainty and were likely or possibly misclassified was actually lower in March 2002. Passing a student who is very slightly below the cutoff or making a student who is only slightly above the cutoff take the exam again are not seriously unjust outcomes. Passing students significantly below the minimum or making students significantly above the minimum take the exam again are more serious problems. A decline of the frequency of such problems is good news.

TABLE 2.21 Comparison of Accuracy Statistics for 2001 and 2002 Test Forms

Statistic	March 2001	May 2001	March 2002
English-language arts			
Zone of Uncertainty	337–361	338–361	335–365
Percent in this Zone	22.4%	23.7%	34.6%
Total Possible Classification Errors	7.1%	7.4%	8.5%
Errors Outside the Zone	0.8%	0.9%	0.4%
Mathematics			
Zone of Uncertainty	339–359	338–363	336–362
Percent in this Zone	20.2%	19.4%	32.6%
Total Possible Classification Errors	6.5%	6.2%	8.2%
Errors Outside the Zone	0.8%	0.8%	0.6%

Table 2.22 shows estimates of the standard error of measurement for different score levels. Students will score within one standard error of true value about two thirds of the time and within two standard errors of their true value over 95 percent of the time. These estimates will differ slightly from estimates provided by ETS. We used a statistical methodology based on more robust item response theory models that include modeling the effects of guessing. One consequence of our approach is that it is not possible to model true achievement levels below chance guessing, although it is certainly possible to receive a scale score based on worse than average luck in guessing. In Chapter 3, we argue for ignoring scores that are significantly below the guessing level and are not so concerned that we cannot estimate standard errors for these low score levels.

TABLE 2.22 Standard Error of Measurement at Different March 2002 Score Levels

Score Level	Estimated Standard Error of Measurement	
	ELA	Mathematics
300	11.8	9.8
325	10.2	8.6
350	9.8	8.1
375	10.0	8.7
400	11.7	10.4
425	14.0	12.5

Summary

A considerable part of our Year 3 effort involved monitoring and analyzing the development, administration, and scoring of the March 2002 CAHSEE. Three things made this an important process to watch: the change in the test development contractor, the very tight timeline for the assembly of the March 2002 form, and the implementation of revised administration procedures.

Our review of test development focused on the quality of new CAHSEE test questions. Activities included monitoring several different types of item reviews conducted by the development contractor—ranging from traditional reviews by content experts to cognitive laboratories as an alternative way of identifying possible flaws in test questions. We also conducted an independent review of the quality of the test questions.

The processes used by ETS were both thoughtful and thorough. Results from our independent item review workshops suggest that the general quality of the test questions remains high. Nonetheless our panelists did identify a number of types of problems that could limit the validity of some test questions and they raised some issues about specific questions.

We observed workshops that prepared district and school personnel for the 2002 test administrations and observed the March and May administrations at six different sites. We also surveyed testing coordinators from our longitudinal sample of schools about their experiences with the 2002 CAHSEE administration. Finally we analyzed accommodations and modifications used in the March 2002 administration.

We noted a number of improvements in the preparation and logistics for the 2002 administrations and did not observe any significant problems. We do offer some suggestions for future improvements in the process in the text above. While there was still some confusion about accommodations and modifications, procedures were much more clearly specified than they were in the 2001 administration. A significant number of students (nearly 10,000) were given accommodations or modifications. It was particularly noteworthy that over 4,000 students used calculators for the mathematics exam, a modification that invalidated their scores. For the most part, passing rates were still very low for students who were allowed accommodations (and would have been for students receiving modifications, had their scores counted).

We also reviewed the process for training scorers for the essay questions and analyzed the consistency of the scores that resulted from their efforts. ETS' process for scoring the essays was new and innovative. Scoring consistency results from this first application of the process were similar to those in the 2001 administrations. We offer a number of specific suggestions for possible improvements to the training and monitoring of essay scorers in the text above.

We conducted analyses of preliminary data from the March 2002 administration to verify ETS' proposed equating of the March 2002 test form to the base form used in the March 2001 administration. We used a divergent approach to test a number of the assumptions underlying that ETS approach. Our results were highly consistent with the results from the operational equating developed by ETS.

We also examined the accuracy of the March 2002 test form. As in the past, we looked at the accuracy with which scores from this form classified students as passing or not passing. Our results showed that more of the March 2002 examinees were very near the minimum passing level than in either of the 2001 administrations. The March 2002 form had a slightly wider "zone of uncertainty" but we estimated that there were fewer misclassifications of students who were significantly below or above the minimum passing level.